N73-16494

Information Note 021072

A THREE-STAGE SAMPLING MODEL FOR REMOTE SENSING APPLICATIONS

by

Ludwig M. Eisgruber
Laboratory for Applications of Remote Sensing (LARS)
Purdue University
West Lafayette, Indiana

Information Note 021072 LARS/Purdue University

A THREE-STAGE SAMPLING MODEL FOR REMOTE SENSING APPLICATIONS

by

Ludwig M. Eisgruber

PART 1: CONCEPTUAL MODEL

1.1. PROBLEM DEFINITION

Large-scale applications of remote sensing for the purpose of preparing crop estimates, natural resource inventories, disaster assessments, etc. for a given geographic region will, in general, involve questions of sampling, since complete coverage of the total geographic region and subsequent analysis of data collected with complete coverage appear technically and economically infeasible. This is true regardless of whether an aircraft, or a satellite is involved, and it applies equally to photography, multispectral measurements, radar, etc. Thus, even if remote sensing provided completely accurate data, estimates (of crop acreage, natural resources, extent of disaster, etc.) for the total region under study will be subject to an error, the so-called error of estimate. This error arises due to the fact that inferences based on selected observations within the region are drawn regarding the characteristics of the total region.

It is the purpose of this discussion to present a conceptual model (in Part 1) and an empirical application (in Part 2) of the relationship between the manner of selecting observations (i.e. the sampling scheme) and its effect on the precision of estimates (i.e. the magnitude of the error of estimate) from remote sensing. Because of technical and practical considerations, a sampling scheme which suggests itself as being useful is a three-stage sampling scheme. 1/ The first stage in this scheme is flightlines, the second stage is segments within flightlines, and the third is units within segments. In general, it can be expected that the various stages contribute differentially to the error of estimate. Also, the contribution from the various stages to the error of estimate is affected by the number of observations in each of the stages (i.e. the subsampling ratios). For instance, an increase in the number of flightlines to be analyzed may be both costly and difficult to execute but decrease the variance of the overall estimate little. On the other hand,

---

1/ The statistical concepts presented here are not new (cf [1] and [2]). They are merely adapted to the problem of remote sensing applications.

an increase in the number of segments per flightline may increase costs and difficulties of analysis little but may have considerable influence on the precision of the estimate. Thus, an arbitrary mix of number of flightlines, segments within flightlines, and units within segments may result in high costs of operation as well as poor estimates. An understanding of the effect of subsampling ratios on the precision of estimates is, therefore, important for most remote sensing applications, particularly those of large scale.

1.2. PROBLEM CHARACTERISTICS AND ASSUMPTIONS

It is assumed that remote sensing is used to estimate a population characteristic (such as acres of a particular crop) in a well-defined geographic region. The flightlines are assumed to be of equal length. Similarly, segments 2/ within each flightline are of equal size, units within each segment are of equal size, and there is an equal number of units in each segment an an equal number of segments within each flightline. Flightline locations are random within the region, as are segment locations within the flightlines and unit locations within the segments.3/

Finally, if a measurement error is present, it is assumed to be constant and/or is random, normally and independently distributed with a mean zero and a standard deviation of σ_ε.

1.3. THE VARIANCE MODELS

In order to ahcieve our objective of investigating the effect of subsampling ratios on the precision of estimates from remote sensing, it is necessary to develop the variance of the estimate in question. We shall do so for both measurement (continuous) and attribute (binomial) data. But first we shall discuss the question of how the measurement error affects the variance.

2/ A "segment" is a sampling unit of specific size (i.e. 1 mile by 10 miles) within a flightline.

3/ If "ground observations" are used to "train" the computer or photo-interpreter, it is assumed to be given. That is to say, a certain classification accuracy is assumed, and the relationship of the amount of ground truth to training, and the level of training to precision of estimates are not explicitly considered in the statistical model to be presented.

1.3.1. The Measurement Error

In remote sensing measurement errors are encountered due to deficiencies in the measuring device, deficiencies in data analysis, etc. Thus, the variance estimates should include a measurement error component.

Let us assume that the relevant mathematical model for the measurement error present in the system under study is

(1.3.1) $$Y_{i\alpha} = G_i + g_i + e_{i\alpha}$$

Where $Y_{i\alpha}$ = value of item obtained in the αth repetition,

G_i = true value of the item,

g_i = constant bias,

$e_{i\alpha}$ = random component.

Since the system under study is one where each item is measured only once, the error $(g_i + e_{i\alpha})$ can be combined into a single term, $\varepsilon_{i\alpha}$, thus simplifying the model to

(1.3.2) $$Y_{i\alpha} = G_i + \varepsilon_{i\alpha}$$

IF the above model (1.3.2) holds, IF the sample is a random sample, and IF we are dealing with an infinite population, then the variance of estimate (to be developed below) will remain valid although no measurement term appears explicitly in the variance definition. However, if we are dealing with a finite population and the measurement error is not explicitly considered, a biased estimate, approximately equal to σ_ε/N will be the result (where N is the number of members in the population). 4/

In either case, the resulting variance will be the variance for the biased mean.

1.3.2. Variance of Estimate of the Mean for Measurement Data and Three-Stage Sampling

The observation y_{ijk} is assumed to be of the form

$$y_{ijk} = \overline{\overline{\overline{Y}}} + u_i + v_{ij} + w_{ijk}$$

4/ cf. [2], p. 305 ff.

where $\bar{\bar{Y}}$ is the overall mean, u_i represents a component associated with the flightline and is constant for all segments within the flightline. The component v_{ij} represents a variation from segment to segment within the flightline, and w_{ijk} represents a variation from data point to data point within the segment. The variates u_i, v_{ij} and w_{ijk} are assumed independently distributed with mean zero. The variates have variances of S_F, S_S, and S_D, respectively (F for flightline, S for segment, and D for data points). The population to be studied contains a finite number of N_F flightlines, N_S segments within each flightline, and N_D data points within each segment. Finally, a sample of n_F, n_S, and n_D observations are randomly chosen for flightlines, segments, and data points, respectively. Then the variance of the sample mean is 5/.

$$(1.3.3) \qquad V(\bar{\bar{y}}) = \frac{(N_F-n_F)}{N_F}\frac{S_F^2}{n_F} + \frac{(N_sN_F-n_sn_F)}{N_SN_F}\frac{S_s^2}{n_Sn_F} + \frac{(N_DN_SN_F-n_Dn_Sn_F)}{N_DN_SN_F}\frac{S_D^2}{n_Dn_Sn_F}$$

An unbiased estimate of $V(\bar{\bar{y}})$ in (1.3.3) is obtained from the sample as follows:

$$(1.3.4) \qquad v(\bar{\bar{y}}) = \frac{1}{n_Fn_Sn_D}\left[\frac{(N_F-n_F)}{N_F}.s_1^2 + \frac{(N_S-n_S)}{N_S}.\frac{n_F}{N_F}.s_2^2 + \frac{(N_D-n_D)}{N_D}\frac{n_F}{N_F}.\frac{n_S}{N_S}.s_3^2\right]$$

The variances S_1^2, S_2^2, and S_3^2 are computed from the sample as follows:

$$(1.3.5) \qquad s_1^2 = \frac{n_Sn_D\sum_i(\bar{\bar{y}}_i-\bar{\bar{\bar{y}}})^2}{(n_F-1)}$$

$$(1.3.6) \qquad s_2^2 = \frac{n_D\sum_i\sum_j(\bar{y}_{ij}-\bar{\bar{y}}_i)^2}{n_F(n_S-1)}$$

$$(1.3.7) \qquad s_3^2 = \frac{\sum_i\sum_j\sum_k(y_{ijk}-\bar{y}_{ij})^2}{n_Fn_S(n_D-1)}$$

5/ If the values of N_F, N_S, and N_D can be considered infinite, or, alternatively, if the ratios of n_F/N_F, n_S/N_S and n_D/N_D can be considered negligible, the finite population correction factors (f.p.c!s) can be omitted and the expression for the variance of the sample mean will reduce to

$$V(\bar{\bar{y}}) = \frac{S_F^2}{n_F} + \frac{S_S^2}{n_F.n_S} + \frac{S_D^2}{n_Fn_Sn_D}.$$

Where $i = 1, \ldots, n_F$

$j = 1, \ldots, n_S$

$k = 1, \ldots, n_D$

and

$$\bar{y}_{ij} = (\sum_k y_{ijk})/n_D$$

$$\bar{\bar{y}}_{i} = (\sum_j \bar{y}_{ij})/n_S$$

$$\bar{\bar{\bar{y}}} = (\sum_i \bar{\bar{y}}_i)/n_F$$

1.3.3. Variance of Estimate of the Mean for Binomial Data and Three-Stage Sampling

In many remote sensing applications the analysis is such that every unit in the population falls into one of two classes, for example C (=corn) and O (=other). Thus:

Number of units in C		Proportion of units in C in	
Population	Sample	Population	Sample
A	a	P=A/N	p=a/n

By means of a simple device it is possible to apply all of the models and formulas developed above to this situation. Suppose, for the moment, that

we are dealing with a simple random sample and single-stage sampling. Define y_i as 1 if the observation is in C and as 0 if it is in O. For the population we then obtain

(1.3.8) $$Y = \sum_{1}^{N} y_i = A$$

(1.3.9) and $$\overline{Y} = \frac{\sum_{1}^{N} y_i}{N} = \frac{A}{N} = P$$

(1.3.10) $$\overline{y} = \frac{\sum_{1}^{n} y_i}{n} = \frac{a}{n} = p$$

Consequently, the problem of estimating A and P can be regarded as that of estimating the total and mean of a population in which every y_i is either 0 or 1. Thus, we can start out with the usual variance formulas in order to develop the variances for proportions. Without actually developing them 6/, we write for the population

(1.3.11) $$V\ (p) = \frac{N-n}{N-1} \quad \frac{P\ Q}{n} \quad ; \quad Q = (1-P)$$

and for the sample (assuming a finite population)

(1.3.12) $$v\ (p) = \frac{(N-n)}{N} \frac{p\ q}{(n-1)} \quad ; q = (1-p)$$

In order to transform (1.3.5), (1.3.6), and (1.3.7) into formulas which are useful for subsampling for proportions, let us proceed as follows:

Let $a_{ij} = \sum_{k} y_{ijk}$, when y_{ijk} is either

zero or one, depending on whether it falls into O ("other") or C("corn"), then

---

6/ See Cochran [2], p. 32 ff.

$$\left.\begin{array}{ll}
(1.3.13) & \bar{y}_{ij} \Rightarrow \bar{p}_{ij} = a_{ij}/n_D = (\sum_k y_{ijk})\ /n_D\ . \\
(1.3.14) & \bar{\bar{y}}_i \Rightarrow \bar{\bar{p}}_i = (\sum_j \bar{p}_{ij})\ /n_S \\
(1.3.15) & \bar{\bar{\bar{y}}} \Rightarrow \bar{\bar{\bar{p}}} = (\sum_i \bar{\bar{p}}_i)\ /n_F
\end{array}\right\}\ \begin{array}{c}\text{Compare}\\ \text{definitions}\\ \text{immediately}\\ \text{following}\\ \text{1.3.7}\end{array}$$

Then

$$(1.3.16) \qquad s_1^2 = \frac{n_S n_D \sum_i (\bar{\bar{p}}_i = \bar{\bar{\bar{p}}})^2}{(n_F - 1)}$$

$$(1.3.17) \qquad s_2^2 = \frac{n_D \sum_i \sum_j (\bar{p}ij - \bar{\bar{p}}i)^2}{n_F\ (n_S - 1)}$$

$$(1.3.18) \qquad s_3^2 = \frac{n_D}{n_F n_S\ (n_{D-1})} \sum_i \sum_j \bar{p}_{ij}\ \bar{q}_{ij}$$

$$\bar{q}_{ij} = (1 - \bar{p}_{ij})$$

Substituting the above definitions into (1.3.3) - or (1.3.4) - will yield the desired variance, v(p).

1.4 PREDICTION OF THE VARIANCE OF ESTIMATE FOR VARIOUS SUBSAMPLING RATIOS

We not only desire to evaluate the precision of estimates for a given sampling scheme, but we are perhaps even more interested in sampling and sub-sampling ratios which are different from those that have been used hitherto. This information is important for planning future experiments and applications of remote sensing on the <u>same type of</u> population.

From the model in (1.3.2) we can predict the variance of $\bar{\bar{y}}$ for different sampling and subsampling ratios. 7/

Suppose in the initial experiment we had values of n_F, n_S, and n_D, respectively, then the variance was

$$(1.4.1) \qquad V(\bar{\bar{y}}) = \frac{S_F^2}{n_F} + \frac{S_s^2}{n_F n_S} + \frac{S_D^2}{n_F n_S n_D}$$

If these values are changed to n'_F, n'_S, and n'_D, respectively, the variance of the sample mean becomes

$$(1.4.2) \qquad V'(\bar{\bar{y}}) = \frac{S_F^2}{n'_F} + \frac{S_S^2}{n'_F n'_S} + \frac{S_D^2}{n'_F n'_S n'_D}$$

In order to utilize this approach, sample estimates of S_F^2, S_S^2, and S_D^2 are required. These may be obtained from the analysis of variance of the sample data as shown in Table 1.4.1 for measurement data. Each of the variance componenets S_F^2, S_S^2, and S_D^2 can be estimated from its mean square and the one just below. 8/ For example

$$S_S^2 = \frac{S_2^2 - S_D^2}{n_D}$$

7/ In the interest of expediency we shall omit all f.p.c.'s from (1.3.3) whenever it is being used in the following discussion. It should be noted that omission of the f.p.c.'s merely results in more conservative variance estimates.

8/ In practice, variance components may turn out to be negative either because the model employed is not relevant or because of the nature of the sampling distributions of variance components (cf [3] and [5], p. 194 ff).

Table 1.4.1 Analysis of Variance for Three-Stage Sampling.

Source of Variation	Degrees of Freedom	Mean Square	Estimate of -
Between flight lines	(n_F-1)	$s_1^2=\frac{n_S n_D \sum_i (\bar{\bar{y}}_i-\bar{\bar{\bar{y}}})^2}{(n_F-1)}$	$S_D^2+n_D S_S^2+n_S n_D S_F^2$
Between segments within flight lines	$n_F(n_S-1)$	$s_2^2=\frac{n_D \sum_i \sum_j (\bar{y}_{ij}-\bar{\bar{y}}_i)^2}{n_F(n_S-1)}$	$S_D^2+n_D S_S^2$
Between data points within segments	$n_F n_S(n_D-1)$	$s_3^2=\frac{\sum_i \sum_j \sum_k (y_{ijk}-\bar{y}_{ij})}{n_F n_S (n_D-1)}$	S_D^2

While the above discussion utilizes expressions (1.4.1 and 1.4.2) which relate to measurement data, a translation to binomial data can readily be made on the basis of discussion in Section 1.3. The relationships in (1.4.1 and 1.4.2) hold, only the computational procedures changes.

1.5. OPTIMAL SAMPLING AND SUBSAMPLING FRACTIONS

These depend on the relationship expressed in (1.3.3) or (1.3.4), respectively, as well as on the cost function relevant to the system. The following cost function is proposed: 9/

$$C = c_F \cdot n_F + c_S \cdot n_S \cdot n_F + c_D \cdot n_D \cdot n_S \cdot n_F +$$

$$S_D \cdot n_D \cdot n_S \cdot n_F + r_D \cdot n_D \cdot n_S \cdot n_F \tag{1.5.1}$$

9/ This is a highly simplified cost function and should be considered as being illustrative only.

Where

C = total cost (\$) of collecting and analyzing data

c_F = cost of flying a flightline of a given length and width

c_S = cost of collecting data over a segment of a given length and width

c_D = cost of analyzing a data point of a given length and width

S_D = cost of storing the (analyzed) data point

r_D = cost of retrieving the results from a data point.

For a given authorized total cost ($\equiv$ available budget) we desire to select values for n_F, n_S, and n_D such that $V(\bar{\bar{\bar{y}}})$ (or $V(\bar{\bar{\bar{p}}})$ is minimum. This is a problem of constrained minimization, and we shall write

$$(1.5.2) \qquad V(\bar{\bar{\bar{y}}}) + \lambda\ (C - c_F . n_F - \ldots - r_D . n_D . n_S . n_F) = 0$$

where λ = Lagrangian multiplier

or, substituting (1.3.3) into (1.5.2),

$$(1.5.3) \qquad \frac{(N_F - n_F)}{N_F} \cdot \frac{S_F^2}{n_F} + \ldots + \frac{N_D N_S N_F - n_D . n_S . n_F}{N_D N_S N_F} \cdot \frac{S_D^2}{n_D n_S n_F} + \lambda\ (C - \ldots .) = 0$$

Differentiating (1.5.3) with respect to n_F, n_S, and n_D, respectively, and setting the resulting equations equal to zero will result in a set of three equations which, when solved, will yield the optimum values in $n_{F_{opt}}$, $n_{S_{opt}}$, and $n_{D_{opt}}$. Sample estimates for S_F^2, S_S^2, and S_D^2 will have to be used. Their computation is discussed in Section 1.3.

By solving the set of equations resulting from a differentiation of (1.5.3) repeatedly for different values of C, a performance function may be traced out, showing the relationship between the magnitude of the variance and an ever costlier data collection scheme. It is hypothesized that this relationship will have the general form of a hyperbola (See Figure 1.4.1). The area above the performance function (in Figure 1.4.1) may be termed the "irrational region", since an improvement can always be achieved for a situation such as represented by point A in Figure 1.4.1, for a given cost, C, be rearranging the subsampling ratio so that a movement out of the "irrational region" onto the performance function occurs. The result will either be a smaller sampling error, $V(\bar{\bar{\bar{y}}})$, for a given cost, C, (a downward movement onto the performance function) or a lower cost, C, for a given size sampling error, $V(\bar{\bar{\bar{y}}})$, (a leftward movement onto the performance function.

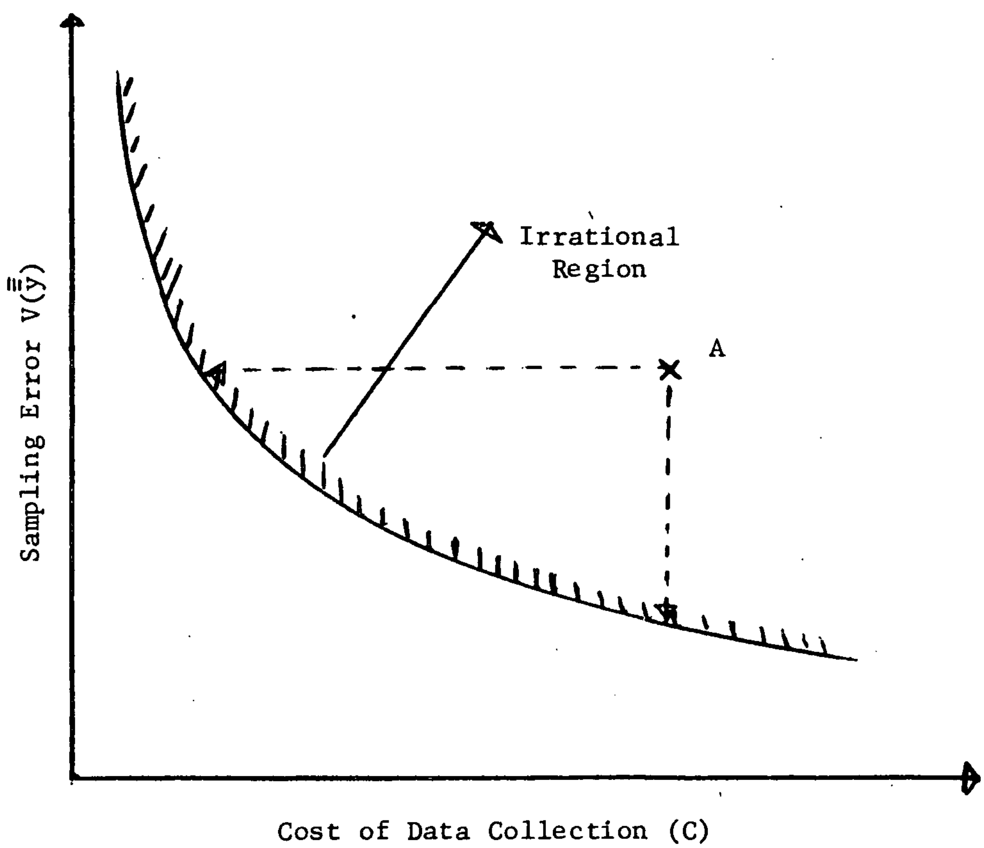

Irrational
Region
A
Sampling Error $V(\bar{\bar{y}})$
Cost of Data Collection (C)

Figure 1.4.1

PART 2. EMPIRICAL ESTIMATES

2.1. OBJECTIVES

In Part II of this paper, an empirical evaluation of the precision of remote sensing estimates of the "acreage of corn in a given region" will be developed. The effect of various subsampling ratios on the precision of estimates will also be investigated empirically.

2.2. THE DATA

2.2.1. Site of the Experiment

The site of the experiment from which the data are taken is that of the "Intensive Study Area" of the 1971 Corn Blight Watch Experiment (CBWE). This area is comprised of the western-most portion of the state of Indiana, a region which is approximately forty (40) miles wide (in an east-west direction) and extending over the entire (north-south) length of the state (see Figure 2.2.1).

2.2.2. Source of Data

The data used for deriving empirical variances of estimate of corn acreage are the multi-spectral scanner data 10/ collected on Mission 43 M of CBWE. (See Appendix Table and also Table 4, Appendix E, Multi-spectral Data Reliability Analysis, [4]. These data were collected over thirty (30) randomly selected segments. Each of these segments was approximately 1 mile wide and 10 miles long. Data for all segments were collected with identical instruments and identical techniques. However, data collected over fifteen of the segments were analyzed by the University of Michigan and its data analysis techniques. The other fifteen segments were analyzed by LARS and its data analysis techniques. The location of the "Michigan Segments" and "Purdue Segments" is shown in Figure 2.2.1.

10/ Photographic data are also available for this site and could have been used.

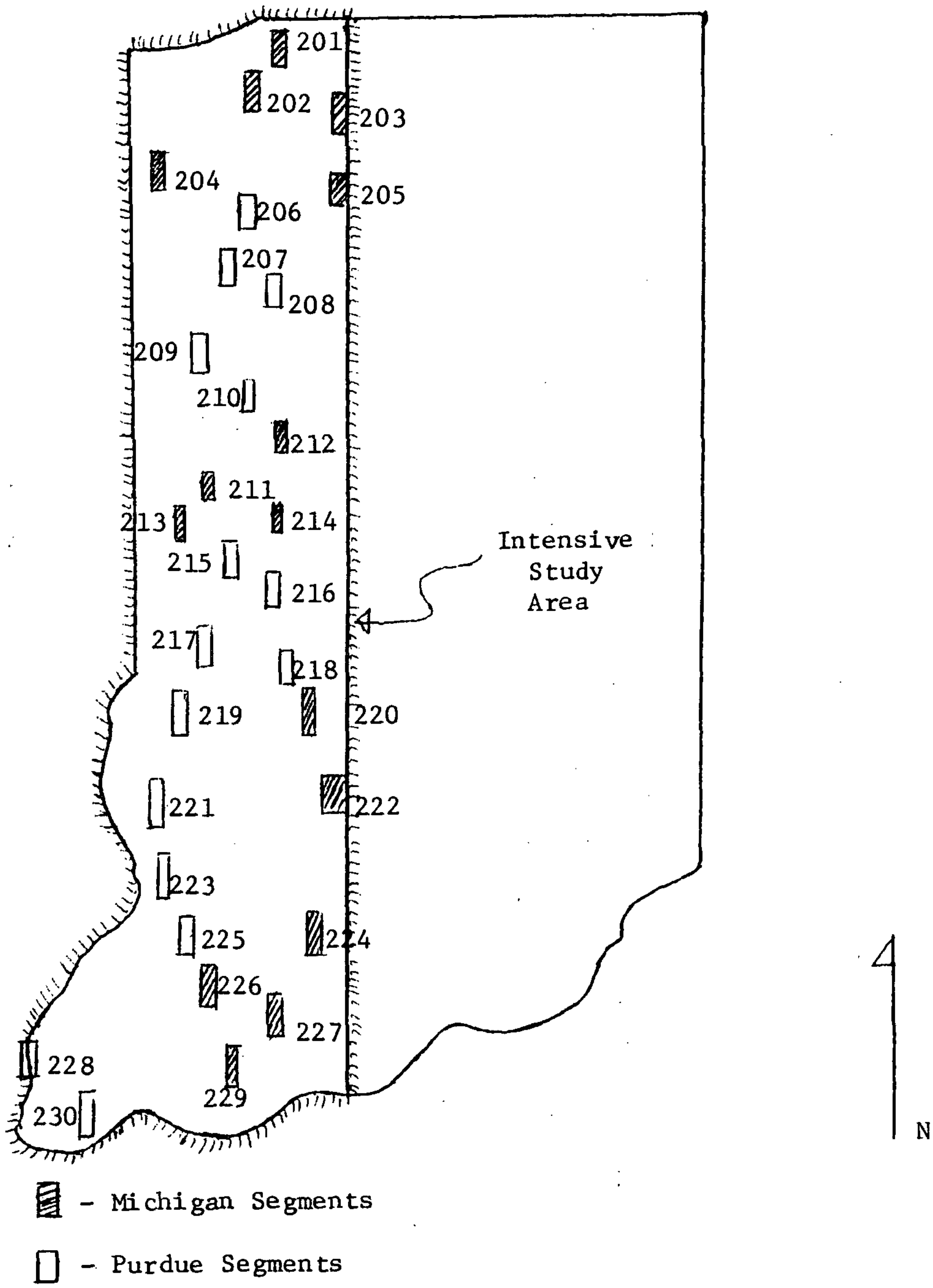

201
202
203
204
205
206
207
208
209
210
212
211
213
214
215
216
217
218
219
220
221
222
223
225
224
226
227
228
229
230
Intensive
Study
Area
N
- Michigan Segments
- Purdue Segments

Figure 2.2.1

2.3. EVALUATION OF DATA

2.3.1. Editing of Data

The data were first examined for consistency. As a result, segments 210, 226, and 228 were eliminated from further analysis.

(a) Segment 210 was eliminated because its area of 7.2 square miles was considerably smaller than the planned 10 square miles for each segment.

(b) Segment 226 was eliminated because its area of 14 square miles was considerably larger than the planned 10 square miles for each segment.

(c) Segment 228 was eliminated because its "planimetered acres" were considerably lower than those for other segments which had a smaller stated "segment area" (in sq. miles), an obvious inconsistency. (Further examination of this particular segment revealed that the segment was an island in the Wabash River.)

2.3.2. Testing for Differences between Michigan Segments and Purdue Segments

The original intent was to utilize data from the 27 segments (30 minus those three eliminated because of inconsistencies). However, because of the differences in analysis techniques there was reason to hypothesize that the data from the Michigan and the Purdue segments have to be viewed as coming from different populations. Therefore, it was necessary to perform appropriate tests before pooling the Michigan segments with the Purdue segments. This was accomplished by testing independently for differences in proportions and differences in variances between the Michigan and the Purdue segments 11/

(a) Difference between Variances of Estimate: The hypothesis tested was

$$H_o : \sigma_M = \sigma_P$$

$$\alpha = .05$$

$$F_{(1-.5\alpha)\ (12,13)} = 1/F_{(.5\alpha)\ (13,12)} = 1/3.569 = .280$$

$$F_{(.5\alpha)\ (12,13)} = 3.525$$

11/ This represents a relatively weak test. However, as will be seen below, the test did distinguish between the two sets of data. Thus, a more complex and powerful test would have added little for the purpose at hand.

where σ^2_M = variance of estimate for Michigan segments

σ^2_P = variance of estimate for Purdue segments

v_M = sample estimate of σ^2_M = .000283

v_P = sample estimate of σ^2_P = .000721

Therefore,

$$F = \sigma^2_P/\sigma^2_M = .000721/.000283 = 2.547$$

Since

$$F_{(1-.5\alpha)} < F < F_{(.5\alpha)}$$

it is not possible to reject $H_o: \sigma_M = \sigma_P$.

(b) <u>Difference between Proportions</u>: The hypothesis tested was

$$H_o: \Pi_M = \Pi_P$$

$$\alpha = .05$$

$$z(.5\alpha) = \pm 1.96$$

Where Π_M = proportion of area in corn in Michigan segments

Π_P = proportion of area in corn in Purdue segments

$\bar{P}_M$ = sample estimage of Π_M = .1514

$\bar{P}_P$ = sample estimate of Π_P = .2900

Therefore,

$$z = \frac{\bar{P}_M - \bar{P}_P}{\sigma_{(\bar{P}_M - \bar{P}_P)}} = \frac{.1514 - .2900}{.0009} = -\frac{.1386}{.0009} = -462.0$$

where $$\sigma_{(\bar{P}_M - \bar{P}_P)} = \sqrt{\Pi(1-\Pi)\left(\frac{1}{n_M} + \frac{1}{n_P}\right)} = .0003$$

$$\text{and} \quad \Pi = \frac{\sum\limits_{mi}\sum y_{im} + \sum\limits_{pj}\sum y_{jp}}{n_M + n_P} = \frac{1,502,444}{7,268,439} = .2067$$

where

n_M = number of data points in Michigan segments

n_P = number of data points in Purdue segments

Y_{im} = value of ith observation in mth Michigan segment

Y_{jp} = value of jth observation in pth Purdue segment

Since

$$Z > Z_{(.5\alpha)}$$

the hypothesis H_o: $\Pi_M = \Pi_P$ must be rejected.

2.3.3. Further Examination of the Difference between Michigan and Purdue Segments

Rejection of the H_o: $\Pi_M = \Pi_P$ necessitates the conclusion that the multi-spectral data from the Michigan and the Purdue segments may not be pooled for purpose of this analysis. However, before proceeding with separate analysis for either the Michigan or Purdue segments, it is important to examine whether Π_M differs from Π_P because of differences in analysis techniques or because of true differences in the proportion of land in corn in the areas where the two sets of segments were located. If the latter is the cause for the difference, then neither set of segments alone is useful for producing estimates for the entire Intensive Study Area.

To examine this question, "ground observations" for each set of segments were compared to each other as well as to the multi-spectral data of the respective set of segments. While no formal statistical tests were made, data in Table 2.3.1 indicate that estimates from "ground observations" agree well with estimates from multi-spectral data for the Purdue segments. However, a substantial downward bias appears to be present in the multi-spectral data for the Michigan segments. Therefore, only Purdue segments will continue to be used in the following analysis.

Table 2.3.1. Comparison of Estimates from Ground Observations with Estimates from Multi-spectral Scanner (MSS) Data for the Michigan and Purdue Segments

Source of Estimate	Michigan Segments $\bar{P}$	$v(\bar{p})$	Confidence Interval*	cv(%)	Purdue Segments $\bar{P}$	$v(\bar{p})$	Confidence Interval*	cv(%)
Ground Observations	.2377	.001108	-	14	.2745	.001739	-	15
MSS Data	.1514	.000283	.1159 -.1875	11	.2900	.000721	.2328-.3472	9

$*\bar{P} \pm t_{.05}\sqrt{v(\bar{p})}$

2.4. THE VARIANCE OF THE ESTIMATE

2.4.1. Delineation of Flightlines

The segments in the Intensive Study Area were not selected on a flightline basis. Instead, they were selected on a random basis. In order to permit an analysis of the effect of different subsampling ratios on the precision of the estimate from a three-stage sampling scheme (flightlines, segments within flightlines, and data points within segments), hypothetical flightlines had to be constructed from the available data. Such construction of hypothetical flightlines assumes that "movement" of segments onto flightlines will not destroy the validity of the data.

Figure 2.4.1 shows that three (hypothetical) flightlines were used. This figure also shows the necessary "movement" of each of the four segments per flightline into the respective flightlines.

2.4.2. Computation of the Variance

The computation of the variance of estimate and the variance components for each of the stages follows the procedure which is described elsewhere (see [1]). 12/ The results are summarized in Table 2.4.1.

12/ Minor modifications were made to account for variability in the number of data points per segment.

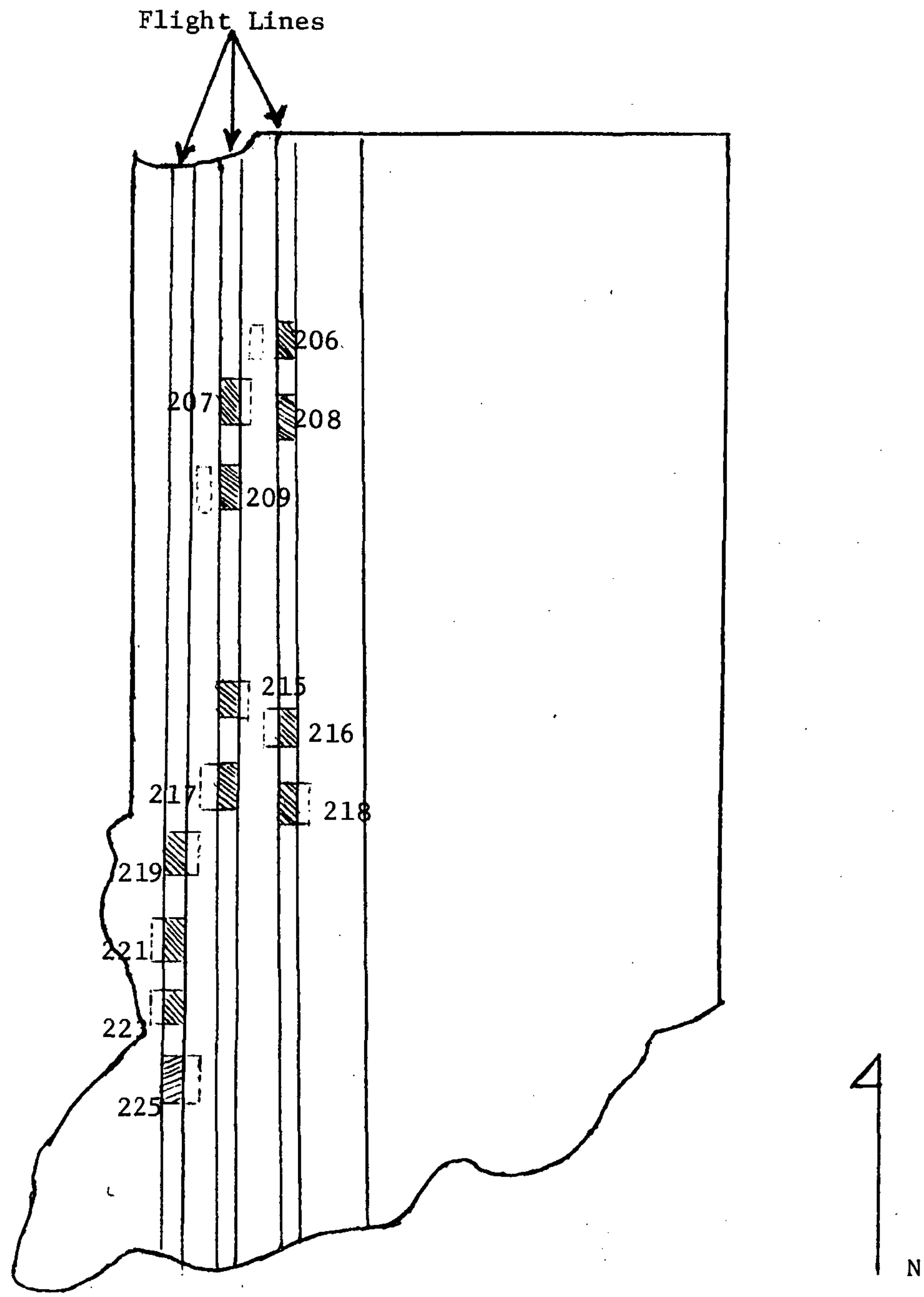


Figure 2.4.1

Table 2.4.1. Analysis of Variance for a Three-Stage Sampling Scheme in Remote Sensing (Three Flightlines - $n_F = 3$; four segments per flightline $-n_S = 4$; 199,675 data points per segment $-n_D = 199,675$).

Source of Variation	Degrees of Freedom	Mean Squares (ns's)	Estimates of -
Between flightlines	2	1,558.5	$S_D^2 + n_D S_S^2 + n_S n_D S_F^2$
Between segments within flightlines	9	3,726.6	$S_D^2 + n_D S_S^2$
Between data points within segments	2,396,101	.2005	S_D^2

In this experiment the f.p.c. cannot be ignored. Therefore, the variance of estimate follows directly from (1.3.4) and (1.3.13)-(1.3.18). Given $N_F = 44$, $N_S = 26$, and $N_D = 31,948 \times 10^3$, then the sample value of the variance of estimate is 13/

$$v(p) = .0006957.$$

The variance components S_F^2, S_S^2, and S_D^2 can now be estimated from each mean square and the one just below. However, S_F^2 turned out to be negative. 14/ If it can be assumed that observations within flightlines are random samples from a normal population, then a test on the intraclass correlation coefficient, $H_O: p_I = 0$ becomes equivalent to $H_o: S_F^2 = 0$. Such a test was executed as follows (cf. [5], p. 194ff.):

$$H_o: S_F^2 = 0$$

$$\alpha = .05$$

$$F_{.95(2,9)} = 4.26$$

$$F = \frac{MS_F}{MS_S} + \frac{1,558.5}{3,726.6} = .4182$$

13/ Had the f.p.c. been ignored, $v'(\hat{p}) = .0006504$. Compare this to $v(p) = .00072$ (Table 2.3.1) where v(p) was computed under the assumption of a simple random sample and where the f.p.c. was ignored.

14/ This "is not only possible but likely in a design such as this." See [3].

Since $F < F_{.95(2,9)}$, H_0 cannot be rejected. Therefore, the variance components utilized in the subsequent analysis are as follows:

$$S_F^2 = 0$$

$$S_S^2 = .0187$$

$$S_D^2 = .2005$$

2.5. EFFECT OF SUBSAMPLING RATIOS ON PRECISION OF ESTIMATES

The formula (2.5.1) was evaluated for various values of n_F, n_S, and n_D. The results are shown in Figures 2.5.1-2.5.3. 15/

$$(2.5.1) \quad v(p)=\left(\frac{1}{n_F} - \frac{1}{N_F}\right) S_F^2 + \left(\frac{1}{n_F n_S} - \frac{1}{N_F N_S}\right) S_s^2 + \left(\frac{1}{n_F n_S n_D} - \frac{1}{N_F N_S N_D}\right) S_D^2$$

Perhaps the most striking observation is that collection and analysis of a large number of data points within segments does not improve the precision of estimate in this particular application. While on the average nearly 200,000 data points were actually analyzed in the experiment, our calculation shows that this did not improve the precision of estimate over that which is derived from $n'_D = 50,000$, given certain values of n'_F and n'_S. Indications are that a considerably smaller number of observations within segments would be satisfactory (see Figure 2.5.1).

Because S_F^2 turned out to be zero in this analysis, the graphs in Figures 2.5.2 and 2.5.3 are merely mirror images. However, both graphs show that the gain in precision of estimates levels off relatively quickly, and the collection of even more data - unless without cost - is likely to become uneconomical rapidly.

15/ For an explanation of the underlying rational and a definition of variables see Part 1 of this paper, in particular equations 1.4.1 and 1.4.2.

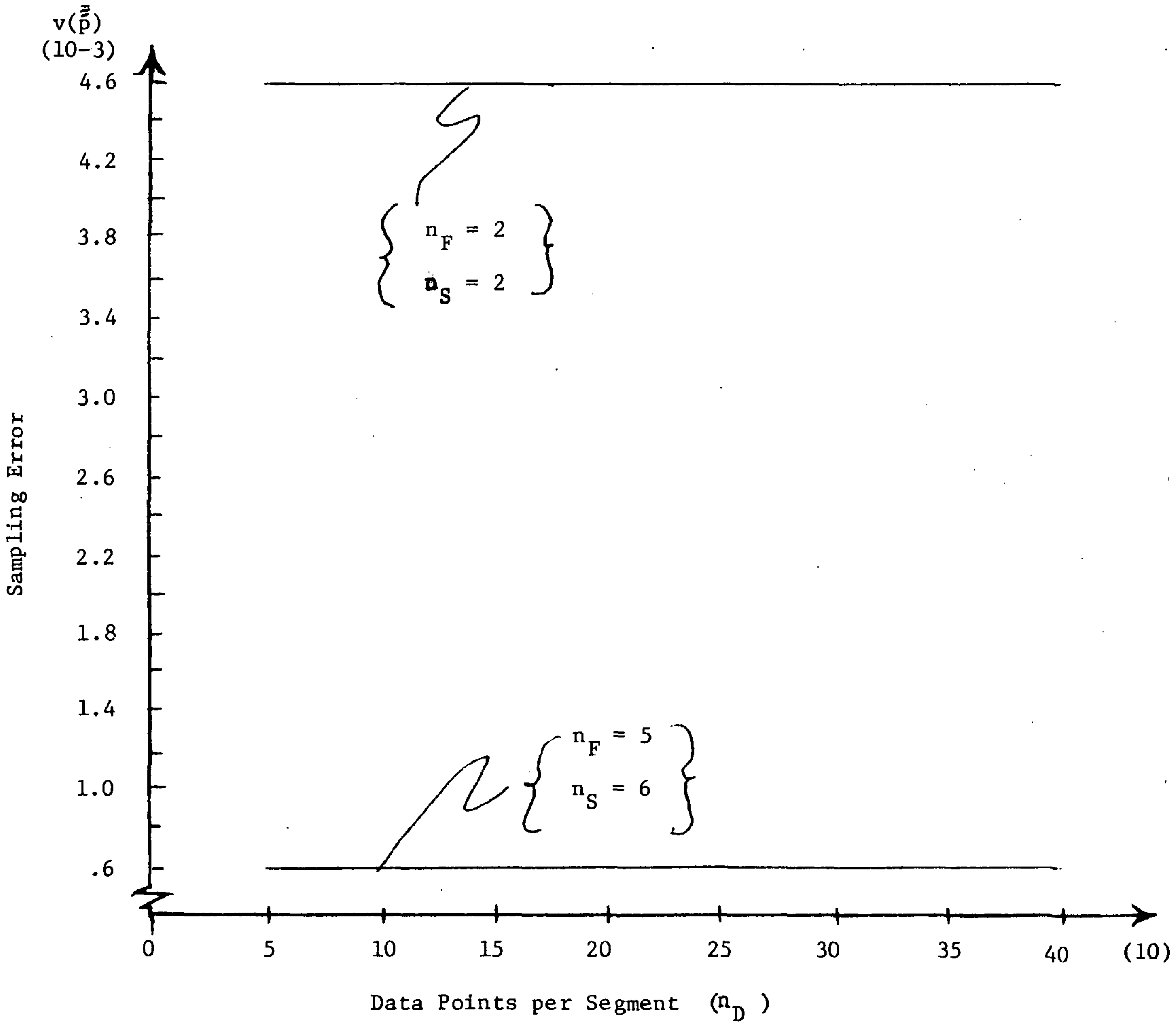
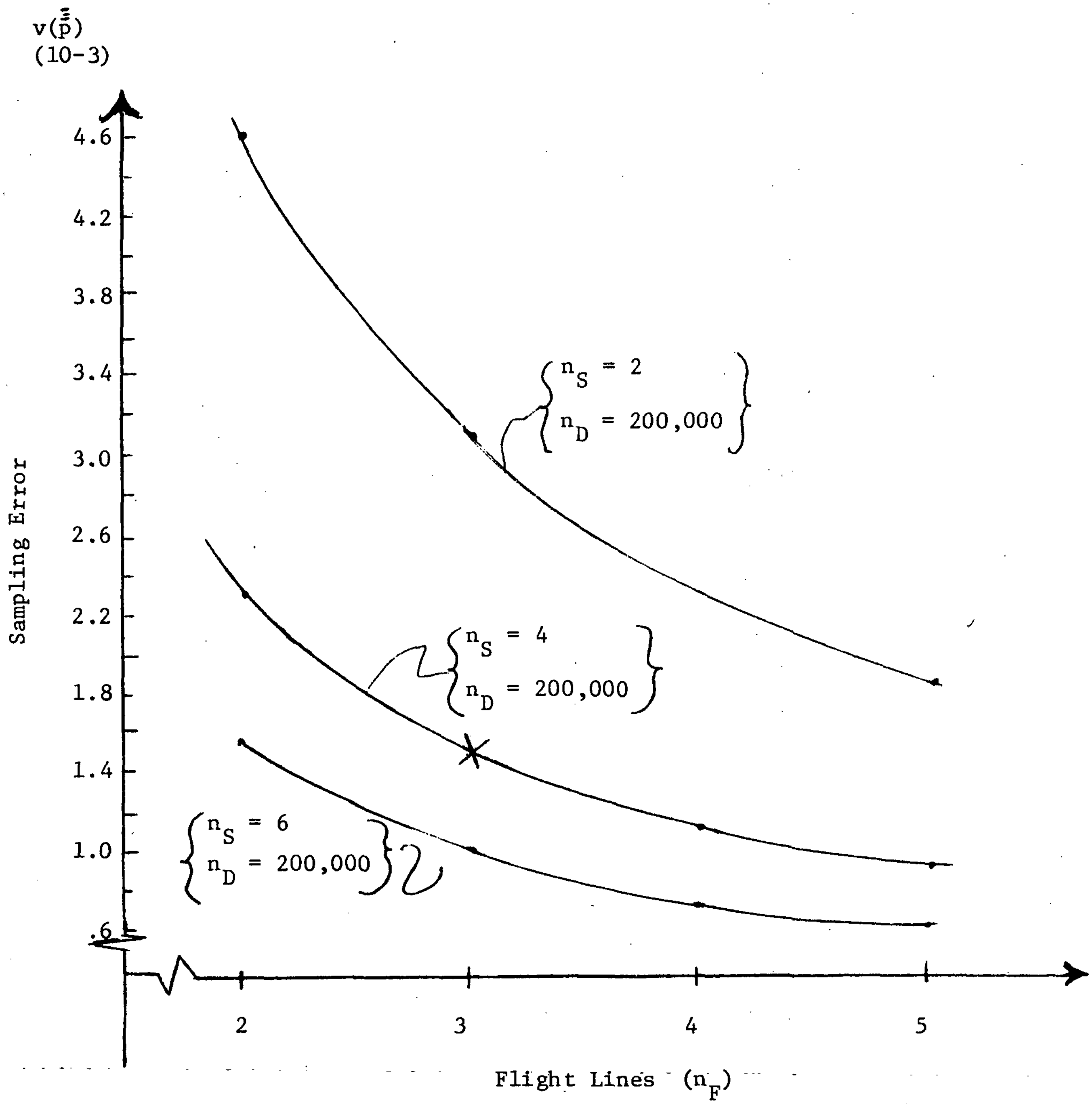

v($\bar{\bar{p}}$)
(10-3)
4.6
4.2
3.8
3.4
3.0
2.6
2.2
1.8
1.4
1.0
.6
Sampling Error
$n_F = 2$
$n_S = 2$
$n_F = 5$
$n_S = 6$
0
5
10
15
20
25
30
35
40
$(10)^4$
Data Points per Segment (n_D)

Figure 2.5.1



v(p̄)
(10-3)
Sampling Error
4.6
4.2
3.8
3.4
3.0
2.6
2.2
1.8
1.4
1.0
.6
$n_S = 2$
$n_D = 200,000$
$n_S = 4$
$n_D = 200,000$
$n_S = 6$
$n_D = 200,000$
2
3
4
5
Flight Lines (n_F)

Figure 2.5.2

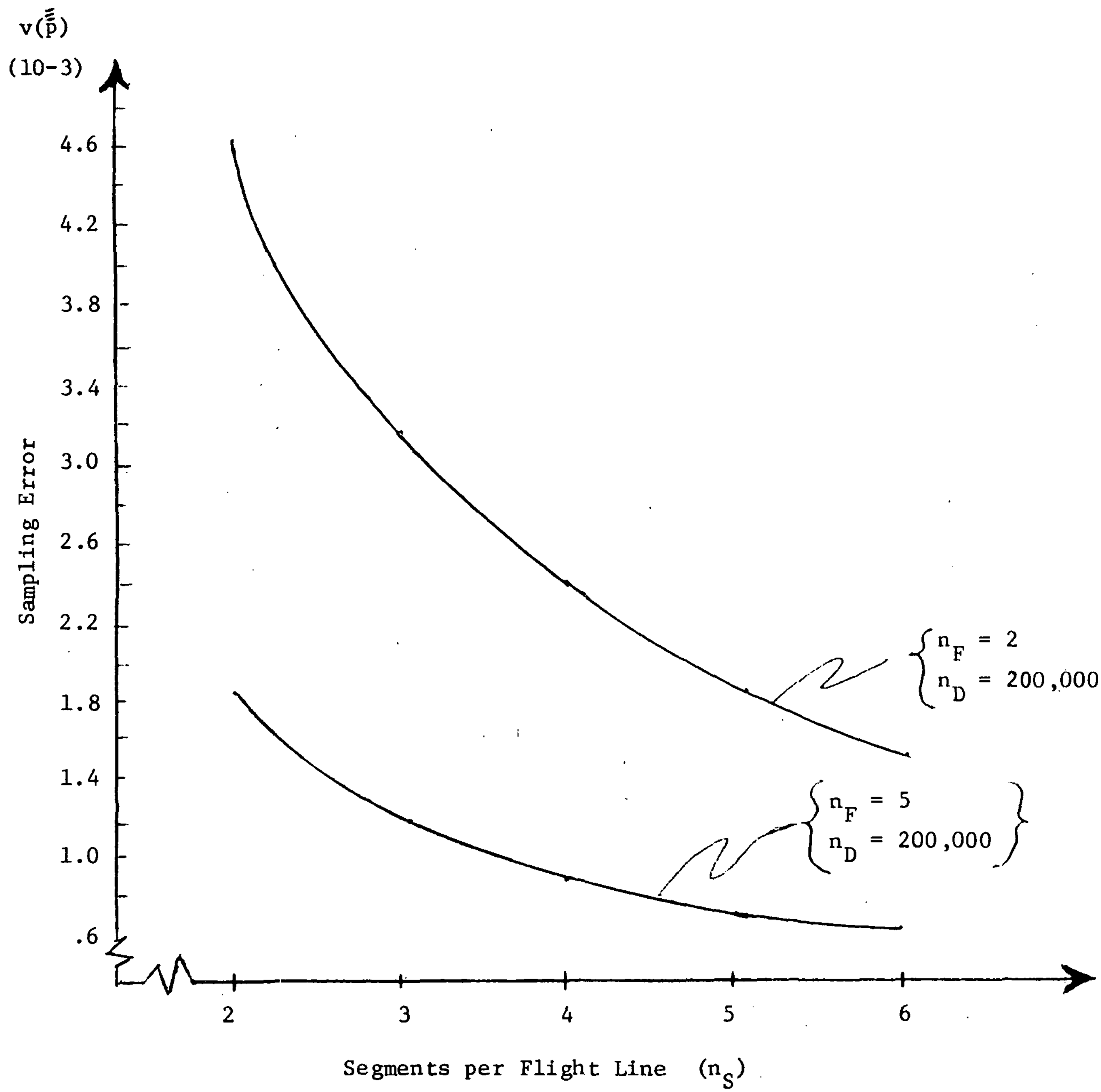

v($\bar{\bar{\bar{p}}}$)
(10-3)
Sampling Error
4.6
4.2
3.8
3.4
3.0
2.6
2.2
1.8
1.4
1.0
.6
2
3
4
5
6
Segments per Flight Line (n_S)
n_F = 2
n_D = 200,000
n_F = 5
n_D = 200,000

Figure 2.5.3

2.6. CONCLUSIONS AND IMPLICATIONS

Results from this study are not nearly as important because of what they show regarding the precision of estimate for Mission 43M as they are because of what they suggest as required analyses in order to assure future practical and economical applications of remote sensing. Some of these requirements are as follows:

(1) The statistical theory and model employed here are of the rather standard variety (only insignificantly modified for the application at hand) and make certain assumptions about the measurement error involved. These assumptions are to date untested and may or may not hold. Even if they hold, the results obtained here are at best an unbiased variance about the biased mean. Furthermore, the distribution of variance components in multistage sampling applications to remote sensing needs further study. The fact that in this study the hypothesis $S_F = 0$ could not be rejected does not rule out the possibility that the computed value for S_F was negative because of an irrelevant statistical model.

(2) This study, in not permitting rejection of the hypothesis that $S_F = 0$, points out that we need to develop organized approaches to the use of a priori information. In retrospect, it appears obvious that, given the cropping pattern in the westernmost 44 mile wide strip of Indiana, the collection of sample data over 12 segments in one flightline should yield an estimate as precise as that obtained by collecting data over 4 segments in each of 3 flightlines. But how can this be determined prior to the experiment? It is actively possible that the appropriate use of a priori information (e.g. census data) could provide the needed insights and basis for designing more efficient experiments. Perhaps an approach similar to the one used in "Project Chitter [6] would be fruitful.

(3) Subsampling ratios and their effect on precision of estimates need to be examined. This study points out strikingly that there is the temptation to oversample in some stages without resulting gains in precision (albeit with increasing costs of data storage and analysis).

(4) To date we know nothing about the relationship between costs, subsampling ratios, and precision of estimates. Yet, it would seem less costly to collect data over twelve segments in one flightline than to collect data over four segments in each of three flightlines. But how much less costly, and what is the trade-off in precision?

(5) Similarly, we know little about the technical and physical difficulty of collecting data in various ways. How much easier is it to collect ground truth on one flightline versus several flightlines? How much easier is it to collect "good" data over one flightline versus several? Given the presence of a broken cloud cover, what is the effect on the quality of data from a few large segments versus a large number of small segments?

(6) It is not possible to generalize from the results of this study to other applications. Instead, similar analyses are required for other types of applications (eg., estimation of acres in corn at different times during the season, estimation of acres in other crops, estimation of degree of insect and disease infestation).

(7) It is unlikely to be practical to develop a unique sampling scheme for each application. Instead, various applications may have to be viewed in terms of joint costs and joint products. Existing theories of joint costs and joint products and associated optimization procedures should be explored for their relevance.

(8) If resources are limited (as they always are), allocation of resources over time for taking samples (i.e. what time periods reflect important change) must also become an integral part of the analysis. For instance, changes over time in corn blight levels would, in all likelihood, affect the variance and the optimal sampling scheme. On the other hand, "acres in corn" may not be affected by passage of time between planting and harvesting.

(9) When remote sensing is done by aircraft, a sampling scheme such as the three-stage sampling scheme used in this analysis appears useful. However, there is no _a priori_ reason why the same model should hold for remote sensing by satellite, when the satellite sequentially covers the entire region to be studied. Perhaps a simple random sample is more appropriate under such circumstances. Also, when time of overflight can no longer be controlled, the question of the extent to which a broken cloud cover can be used as the sample selection device becomes an interesting and important one.

References

[1] Anderson, R. L., and T. A. Bancroft, Statistical Theory in Research, McGraw-Hill, New York, 1952.

[2] Cochran, W. G., Sampling Techniques, Wiley, New York, 1953.

[3] Leone, F. C., and L. S. Nelson, Sampling Distributions of Variance Components, Technometrics, 8, 3, August 1966, pp. 457-468.

[4] LARS, CBWE Interim Report, Lafayette, Indiana, 1971.

[5] Ostle, B., Statistics in Research, Second Printing, Iowa State University Press, Ames, 1956.

[6] Project Chitter (Acre): 1967 Final Report, Mark Systems, Inc., 2999 San Ysidro Way, Santa Clara, California 95051.

APPENDIX

Appendix Table 1. Mission 43M (August 9, 1971,) Multispectral Scanner Data from the Intensive Study Area.

Segment No. (1)	Michigan (M) or Purdue (P) Segment (2)	Points in Segment (3)	Pct. of Segment Classified as Corn (4)	Acres of Corn (Ground Truth) (5)	Planim. Acres of Segment (6)	Segment Area (Sq. miles) (7)
201	M	387890	12.57	1537	7970	12.0
202	M	301087	15.56	2191	6569	10.0
203	M	298075	8.31	2831	6858	11.5
204	M	379556	20.86	2892	7720	11.0
205	M	332032	19.00	1888	7780	12.0
206	P	158885	35.87	2665	5285	9.0
207	P	233153	36.98	3404	7973	12.0
208	P	225342	44.59	3679	7558	12.0
209	P	165708	43.84	2324	6059	9.0
210	P	130511	21.67	1092	4790	7.2
211	M	289830	11.80	2272	6935	10.5
212	M	306366	13.64	2330	7650	9.5
213	M	245800	19.65	1716	6094	10.0
214	M	262840	16.33	1247	5232	9.0
215	P	154467	24.80	864	5750	9.0
216	P	207218	18.31	1278	6932	10.0
217	P	246752	26.37	1758	8030	11.5
218	P	208094	26.02	318	7022	10.5
219	P	181745	14.81	996	5946	9.0
220	M	244795	8.96	97	5774	8.5
221	P	224446	35.64	2362	5835	9.0
222	M	282812	8.42	338	6000	9.0
223	P	194361	20.64	994	6749	8.5
224	M	221671	2.72	201	5120	9.5
225	P	195930	27.42	2125	7275	10.5
226	M	409600	12.80	887	9121	14.0
227	M	264795	23.42	1490	6774	9.5
228	P	91457	60.44	997	3857	8.0
229	M	261700	29.86	1684	5855	8.5
230	P	161521	21.23	871	5535	8.5